Capstone Project - MovieLens

Stevonne Nugent

12/9/2020

Executive Summary

The purpose of this project is to create a movie recommendation system utilizing a subset of the original MovieLens dataset. This subset used for the assessment contained approximately 10 million movie ratings (reviews) for 10,677 movies by 69,878 users, which was further separated 90% and 10% into edx and validation datasets, respectively. The edx dataset was used for model building and testing, while the validation dataset was used to test the accuracy of the final model developed with the aim of attaining a root mean square error (RMSE) of less than 0.86490.

To facilitate the development of the final model (recommendation system), the average effect of the variables associated with each movie, user, genre and movie release year was used. For model development the edx dataset was randomly separated into a train (edxtrain) and test (edxtest) dataset. The edxtrain set was used to build the model and the edxtest set was used to tune the penalty terms. The model that resulted in the lowest RMSE was chosen and then tested on the validation set to assess its accuracy.

Data analysis was undertaken to examine and explore the data to assist in the process of developing the model. A total of 4 predictors were included in the model, namely movieId (b_i); userId (b_j); genres (b_g); and movie year (b_y). A regularization approach was used in the development of the model based on the fact that some predictors had a small sample size, and optimization of the effects of each variable was executed at the each stage of the model development process.

The final model obtained was then applied to the validation dataset to predict the movie rating of a user producing an RMSE of 0.86464, which was below the target of 0.86490.

Section 1: Introduction

Recommendation systems in general use historical ratings of goods and services by their users to make specific recommendations based on rating predictions. These systems become very useful to consumers as a wide variety of goods and services are available for consumption.

In this project, the goal was to build a movie recommendation system that will predict ratings for a set of users. The movie recommender service MovieLens 10M Dataset from Harper and Konstan 2016 was used. The dataset contained 10,000,054 movie ratings for 10,677 movies, from 69,878 users. To assess the accuracy of the final movie recommendation model developed, the MovieLens dataset was divided into an edx and a validation dataset, the former for training and testing the model and the latter represents the final holdout dataset to test the accuracy of the model. The objective was to build a model to predict user rating of

movies in the validation dataset utilizing the edx data set with a root mean square error (RMSE) of less than 0.86490.

To build the recommendation model, the average effect of movie, user, genre, and movie year was estimated. The rest of the paper is organized as follows:

- Section 2 – Data Description, Exploration & Wrangling
- Section 3 – Model Development
- Section 4 – Results
- Section 5 – Conclusion.

Section 2: Data Description, Exploration & Wrangling

Paramount to any data analysis process are describing the data to be used; exploring the data to see trends and getting more details about the variables to see their potential for inclusion into model development; and any data cleaning that may be needed to have the data in a format most suitable for use.

Data Description

The MovieLens dataset used contained 10,000,054 rows and 6 columns with no missing values. The 6 columns were the following variables:

- userId – unique user identification number, there are 69,878 unique users
- movieId – unique movie identification number, there are 10,677 unique movies
- rating – rating given by each user to a specific movie, ratings are on a 5 star scale with 0.5 point increment, i.e. ratings range from 0.5 to 5.0, there are 10 unique rating scores
- timestamp – the date and time a movie was rated measured in seconds, there are 7,096,905 unique time stamps
- title – contains the title of a movie including the year the movie was released, there are 10,676 unique movie titles
- genres – contains a list of pipe-separated genre of each movie, there are 797 unique genre groups.

Table 1 shows the first 6 rows in the MovieLens dataset. Note, the first 6 rows shows information for only 1 user, with userId 1 who rated 6 different movies (based on unique movieId and title) and gave each the maximum rating of 5.

Table 1
First 6 Rows of the MovieLens 10M Dataset

userId	movieId	rating	timestamp	title	genres
1	122	5	838985046	Boomerang (1992)	Comedy|Romance
1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
1	231	5	838983392	Dumb & Dumber (1994)	Comedy
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi

The aim of this project is to build a model with the ability of predicting the rating a user will give a movie, i.e. rating is the outcome variable and the other variables will be probable predictor variables that will assist in predicting the rating. To mitigate against the possibility of overfitting this model, the original MovieLens dataset which contains approximately 10 million movie ratings (reviews) for 10,677 movies by 69,878 users was randomly separated into edx dataset and validation dataset, which represented 90% and 10% of the original dataset, respectively. The former was used to train and tune the model and the latter was used to evaluate the final model, to ensure this was possible all userId and movieId in the MovieLens dataset needed to be in the edx dataset.

Data Wrangling

The edx dataset was examined and transformed to identify possible predictor variables. Once identified the changes made to the edx dataset will be done to the validation dataset prior to evaluating the final model.

Initial data exploration highlighted that the genres were pipe-separated, the title contained the year the movie was released and the timestamp was in seconds since Midnight Coordinated Universal Time January 1, 1970. There was need to transform these variables into a more useful format as the specific genre, the year the movie was released and the year the movie was reviewed could be potential predictor variables. The following changes were made:

1. convert timestamp to human readable date format
2. extract the month and year from the date
3. extract the release year for each movie from the title. Table 2 shows the first 6 rows of select variables that were processed from the edx dataset to make them more useful.

Table 2
First 6 Rows of Select Variables from the Processed edx Dataset

timestamp	title	reviewdate	reviewyear	reviewmonth	movieyear
838985046	Boomerang (1992)	1996-08-02	1996	8	1992
838983525	Net, The (1995)	1996-08-02	1996	8	1995
838983421	Outbreak (1995)	1996-08-02	1996	8	1995
838983392	Stargate (1994)	1996-08-02	1996	8	1994
838983392	Star Trek: Generations (1994)	1996-08-02	1996	8	1994
838984474	Flintstones, The (1994)	1996-08-02	1996	8	1994

The transformed edx dataset was then randomly separated into edxtrain and edxtest datasets, with a respective ratio of 90% and 10%. Model development took place using the edxtrain dataset to build the model and the edxtest dataset to calibrate the model parameters. As with the edx and MovieLens datasets, it was ensured that the edxtrain dataset had all the movies and users as the edx dataset.

Data Exploration

The edxtrain dataset was explored individually and collectively to examine possible relationships, to put together a list of potential predictor variables.

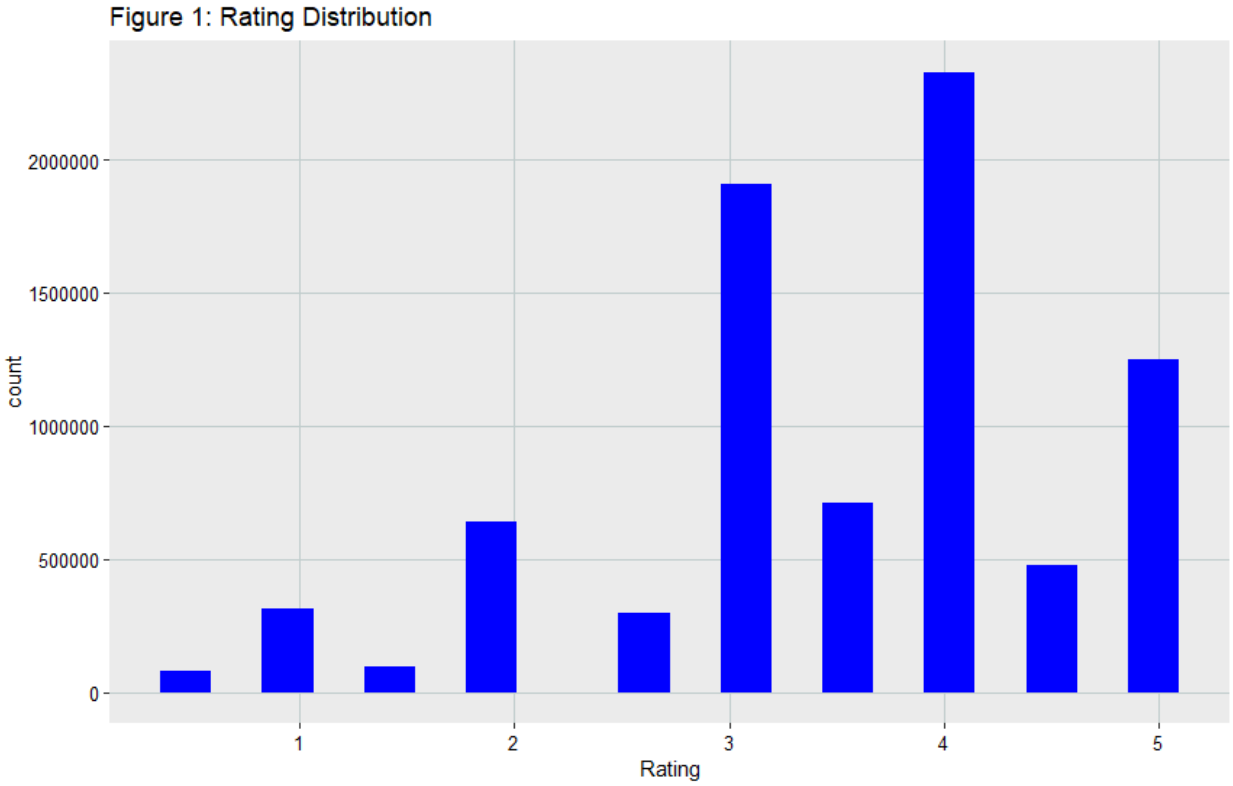
rating

- rating of 'i' movie by 'j' user ranges from 0.5–5.0, with 0.5 point increments (Figure 1)

- the average rating is 3.512 and the median is 4.0 (Table 3)
- rating is more likely to be a whole number than a fraction (see Figure 1).

Table 3
Descriptive Statistics of Ratings in the edxtrain Dataset

Descriptive Stats	Values
minimum	0.500000
maximum	5.000000
first_quartile	3.000000
median_rating	4.000000
third_quartile	4.000000
average	3.512456
std_dev	1.060362
range	4.500000



movieId

- unique identification for each movie
- 10,677 unique movie IDs
- potential predictor variable given the fact that the quality of a movie impacts its rating and the higher the quality the more likely a user will give it a higher rating on average, and vice versa
- dependability of the variable is linked to the number of users that would have given the movie a rating, for example, the average rating for a movie that has been watched many times is more dependable in

terms of how others may rate this movie, compared with an average rating for a movie that has been rated one or two times

- to improve dependability of the rating the number of users that contributed to the movie rating becomes important
- Figure 2 shows that some movies were rated more frequently than others, and Table 4 reveals possible hit movies, for example movieId 593, "Silence of the Lambs" which was the 3rd most rated movie (27,327) with an average rating of 4.203 out of 5
- likewise, less known movies (non-hit movies) were rated less frequently, with 159 movies being rated just once, therefore we would be less confident in using its average rating to predict how other users will rate that movie—for example movieId 53355, "Sun Alley (Sonnenallee)" which was rated once and given a 5-star rating (Table 5).

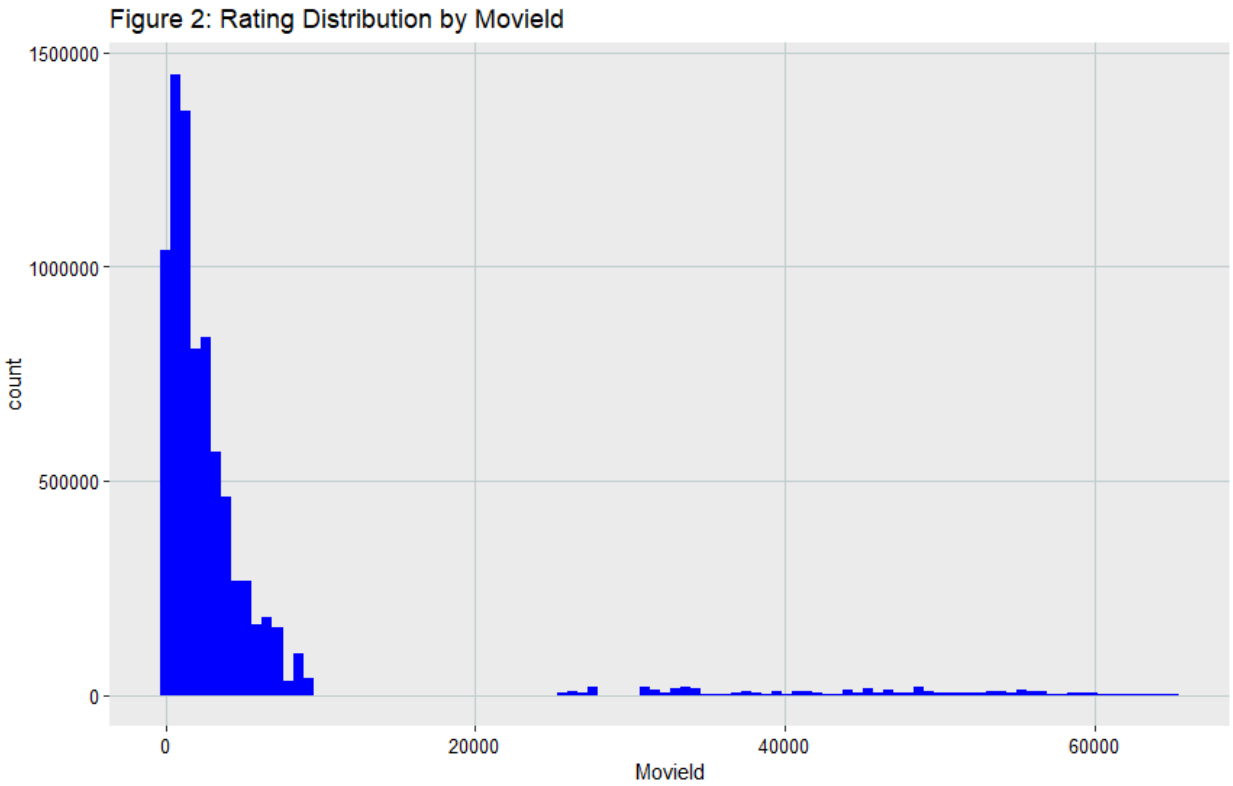


Table 4
Top 10 Rated Movies

movieId	title	count	avg_rating
296	Pulp Fiction (1994)	28168	4.154022
356	Forrest Gump (1994)	27987	4.013328
593	Silence of the Lambs, The (1991)	27327	4.202510
480	Jurassic Park (1993)	26381	3.665574
318	Shawshank Redemption, The (1994)	25188	4.456567
110	Braveheart (1995)	23545	4.081928
457	Fugitive, The (1993)	23397	4.010557
589	Terminator 2: Judgment Day (1991)	23332	3.926346
260	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	23065	4.221028
592	Batman (1989)	21857	3.382692

Table 5
Movies Rated Only Once with High Ratings

movieId	title	rating
53355	Sun Alley (Sonnenallee) (1999)	5.0
51209	Fighting Elegy (Kenka erejii) (1966)	5.0
33264	Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	5.0
42783	Shadows of Forgotten Ancestors (1964)	5.0
3226	Hellhounds on My Trail (1999)	5.0
64275	Blue Light, The (Das Blaue Licht) (1932)	5.0
64418	Man Named Pearl, A (2006)	4.5
7452	Mickey (2003)	4.5
63179	Tokyo! (2008)	4.5
58185	Please Vote for Me (2007)	4.5

To ensure the dependability of the variable movieId in the model and account for low frequency rated movies with high ratings, we will utilize regularization to constrain the variability of size effects by adding a penalty term.

userId

- unique identification for each user
- 69,878 unique users in the edxtrain dataset
- dependability of userId as a predictor variable (user effect) relies on the frequency in which a user rates movies
- as with the movieId variable where some movies were rated more than others, some users are more active than others
- minimum number of ratings by a user was 9, maximum was 5,931 (Table 6a)
- average user effect ranged from 0.5 to 5.0 (Table 6b).

Table 6a
Distribution of Frequency of User Rating

Descriptive Statistics	Number of Reviews
minimum	9.0000
maximum	5931.0000
first_quartile	29.0000
median_rating	56.0000
third_quartile	127.0000
average	115.9172
std_dev	175.5840
range	5922.0000

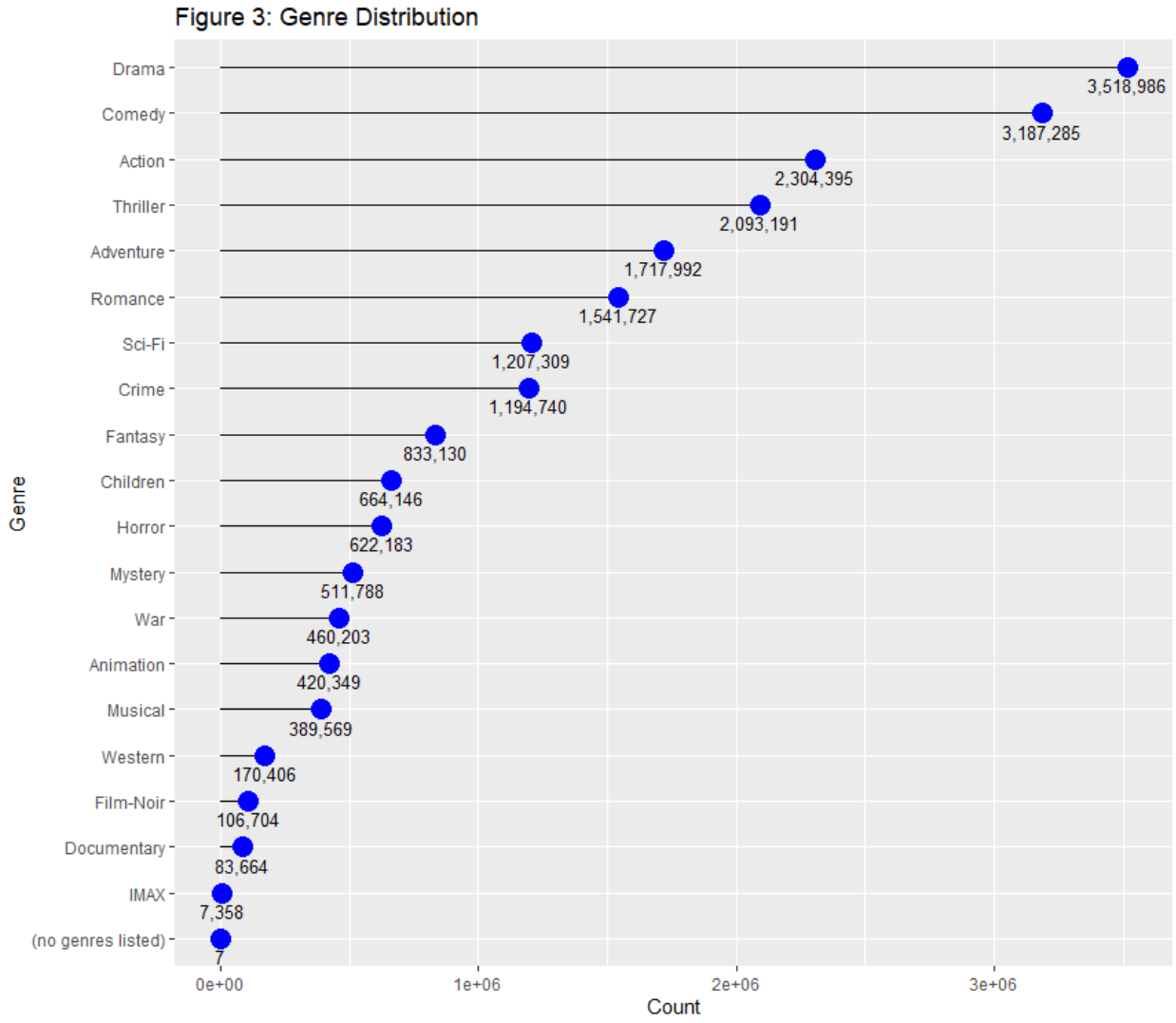
Table 6b
Distribution of Average User Rating

Descriptive Statistics	User Average Rating
minimum	0.5000000
maximum	5.0000000

first_quartile	3.3555556
median_rating	3.6341463
third_quartile	3.9047619
average	3.6135327
std_dev	0.4332773
range	4.5000000

genres

- classified the style/type of movie
- some movies were classified as multiple genres some as single and there are 797 unique genre groups
- there are 18 genres, IMAX (high-resolution movies) and 1 movie with no rating which was watched 7 times (Figure 3)
- top 3 watched genres was Drama, Comedy and Action (see Figure 3), of note these top rated genres (that is, the genre that appears the most individually or as part of a genre group)



- genre grouping with the highest average rating was Animation|IMAX|Sci-Fi at 4.67 (Table 7)
- genre grouping with the lowest average rating was Documentary|Horror at 1.44 (Table 8)
- some genre groups were rated frequently (e.g. Crime|Mystery|Thriller, with 24,173 ratings) {see Table 7} while others were rated infrequently (e.g. Action|Drama|Horror|Sci-Fi, with 4 ratings) {see Table 8}.

Table 7
Top 10 Genre Groupings by Average Rating

genres	count	average
Animation|IMAX|Sci-Fi	6	4.666667
Drama|Film-Noir|Romance	2693	4.303565
Action|Crime|Drama|IMAX	2095	4.299523
Animation|Children|Comedy|Crime	6418	4.278903
Film-Noir|Mystery	5431	4.239275
Crime|Film-Noir|Mystery	3650	4.223973
Film-Noir|Romance|Thriller	2190	4.217352
Crime|Film-Noir|Thriller	4365	4.211455
Crime|Mystery|Thriller	24173	4.201051
Action|Adventure|Comedy|Fantasy|Romance	13329	4.197052

Table 8
Bottom 10 Genre Groupings by Average Rating

genres	count	average
Documentary|Horror	547	1.441499
Action|Horror|Mystery|Thriller	289	1.614187
Comedy|Film-Noir|Thriller	17	1.647059
Action|Drama|Horror|Sci-Fi	4	1.750000
Adventure|Drama|Horror|Sci-Fi|Thriller	196	1.795918
Adventure|Animation|Children|Fantasy|Sci-Fi	627	1.900319
Action|Adventure|Drama|Fantasy|Sci-Fi	51	1.901961
Action|Horror|Mystery|Sci-Fi	19	1.921053
Action|Children|Comedy	460	1.931522
Action|Adventure|Children	745	1.934228

movieyear

- first movie was released in 1915 and the last movie was released in 2008 (Figure 4)
- movies released in 1995 were most rated (see Figure 4)
- movies released in 1946 received the highest average rating of 4.06 (Table 9)
- movies released in 1915 received the lowest average rating of 3.27 (Table 10)

Table 9
Top 10 Movie Years by Average Rating

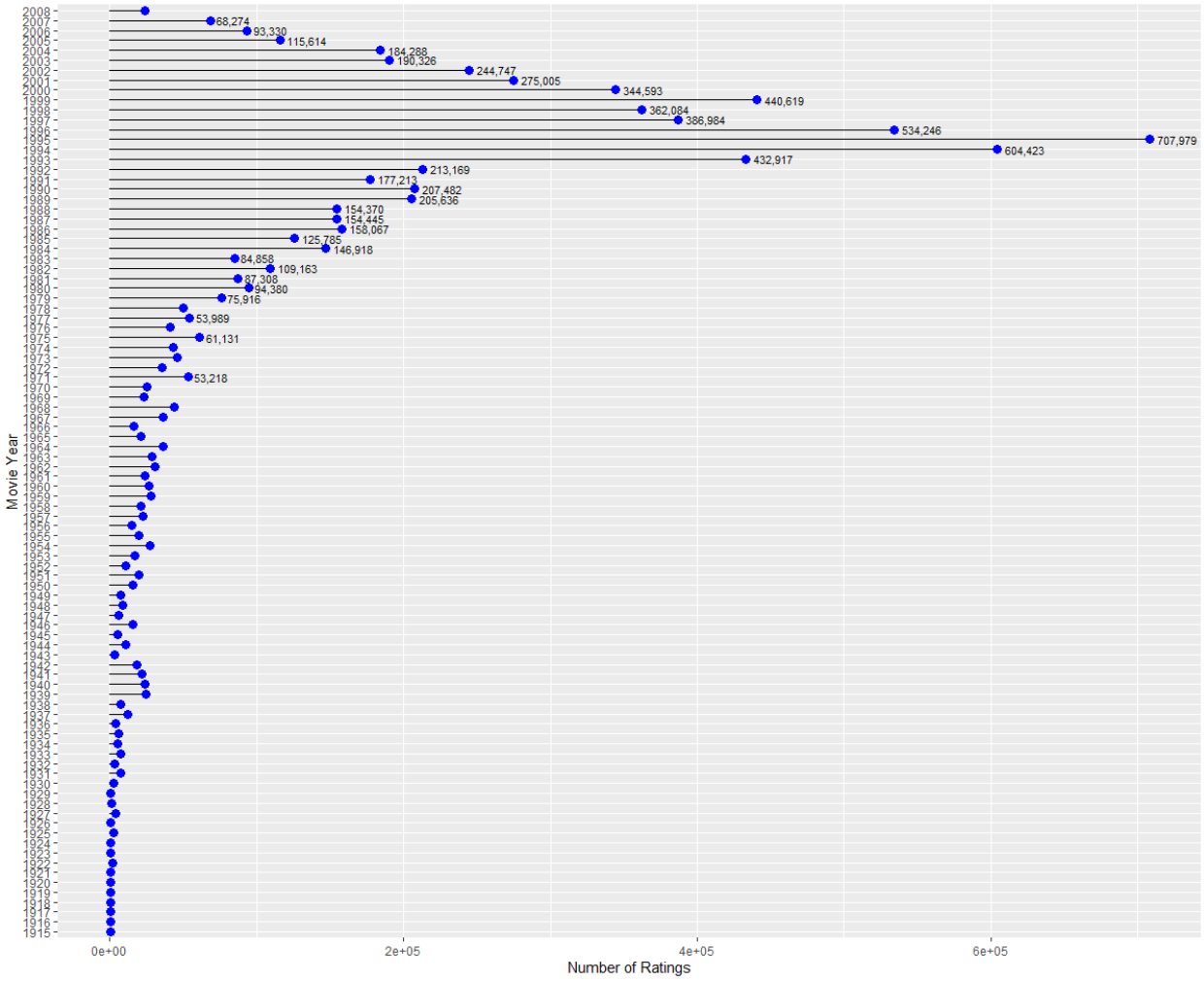
movieyear	average
1946	4.055398
1934	4.050644

1942	4.044690
1931	4.026951
1941	4.021318
1927	4.016492
1954	4.011375
1957	4.009564
1944	3.991993
1962	3.989676

Table 10
Bottom 10 Movie Years by Average Rating

movieyear	average
1915	3.269939
1919	3.284722
1996	3.361101
1997	3.363400
2000	3.394110
1990	3.396051
1998	3.415527
1992	3.431749
2001	3.438150
1995	3.442565

Figure 4: Number of Movie Ratings by Movie Year



Section 3: Model Development

Based on the results obtained in section 2, the 4 variables were used as prediction variables in a phased approach. This was done by adding each variable a step at a time while optimizing the penalty term using the lowest RMSE based on the model to predict ratings in the edxtest dataset for each variable and testing the RMSE at each stage to assess if our target of less than 0.86490 was met. Note that for all models tested the predicted ratings were constrained to a minimum value of 0.5 and a maximum of 5.0, since the ratings ranged from 0.5 and 5.0. The variables used were:

- overall average effect ($\bar{b}$) – average rating of all movies
- movie effect (b_i) – average movie effect while controlling for overall average
- user effect (b_j) – average user effect while controlling for overall average rating and movie effect
- genre effect (b_g) – average genres effect while controlling for overall average rating, movie and user effects
- movieyear effect (b_y) – average movie year effect while controlling for overall average rating, movie, user and genres effects.

The objective is to estimate the rating "r" of each movie "i" by each user "j"—$r_{i,j}$—using the above listed variables.

The estimate for $r_{i,j}$ will be $\hat{b}_{i,j}$ leading to the following model:

$$\hat{b}_{i,j} = \bar{b} + \hat{b}_i + \hat{b}_j + \hat{b}_g + \hat{b}_y$$

After $\hat{b}_{i,j}$ is estimated, its values were constrained by the following conditions:

if $\hat{b}_{i,j} < 0.5$, then $\hat{b}_{i,j} = 0.5$

if $\hat{b}_{i,j} > 5.0$, then $\hat{b}_{i,j} = 5.0$

$$\hat{r}_{i,j} = \bar{b} + \hat{b}_i + \hat{b}_j + \hat{b}_g + \hat{b}_y + \hat{e}_{i,j}$$

$\hat{e}_{i,j}$ is the error term and we will minimize the RMSE

$$RMSE = \sqrt{\frac{1}{N}\sum_{i,j}\left(\hat{b}_{i,j} - r_{i,j}\right)^2}$$

As shown in section 2, some estimates have small sample sizes, for example movies rated only once. To control for this a regularized model was used as follows:

- overall average effect is $\bar{b}$
- movie effect is

$$\hat{b}_i = \frac{1}{n_i + \lambda_i}\sum_{j=1}^{n_j}(r_{i,j} - \bar{b})$$

The sequencing above for the movie effect continues for all other effects user, genre and movieyear each using a different optimal lambda.

Section 4: Results

This section presents the modeling results and discusses the model performance. The model was developed using a phased approach adding each variable one step at a time while checking the RSME. Figure 5 to Figure 8 show the range of lambda values and their respective RMSE for the rating variable in the edxtest dataset. The lambda that minimizes the RMSE is chosen to estimate the value of each effect.

Figure 5: Optimal Value of $\hat{\lambda}_i$ that minimizes the RMSE for $\hat{e}_{ij} = r_{ij} - \bar{b} - \hat{b}_i$

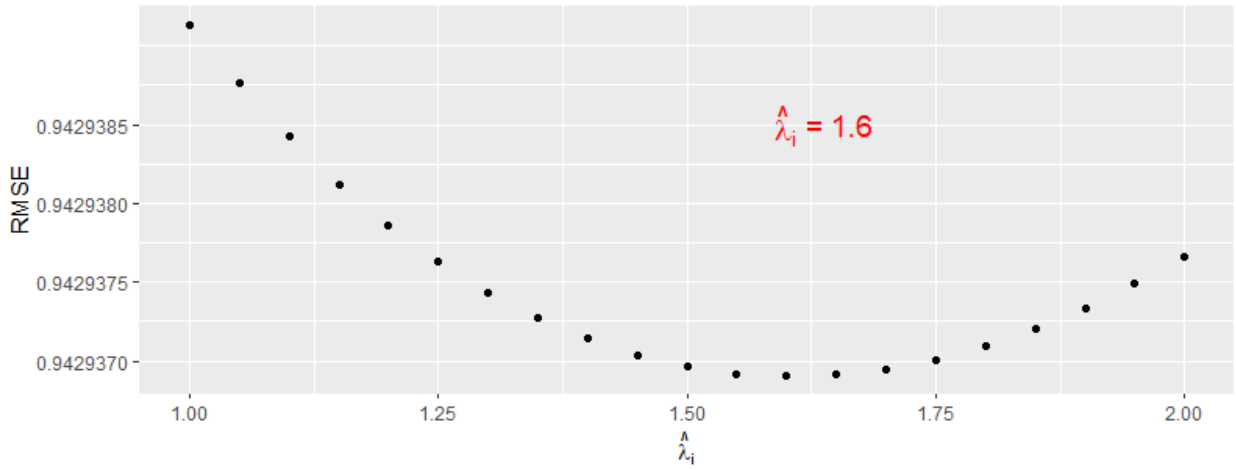


Figure 6: Optimal Value of $\hat{\lambda}_j$ that minimizes the RMSE for $\hat{e}_{ij} = r_{ij} - \bar{b} - \hat{b}_i - \hat{b}_j$

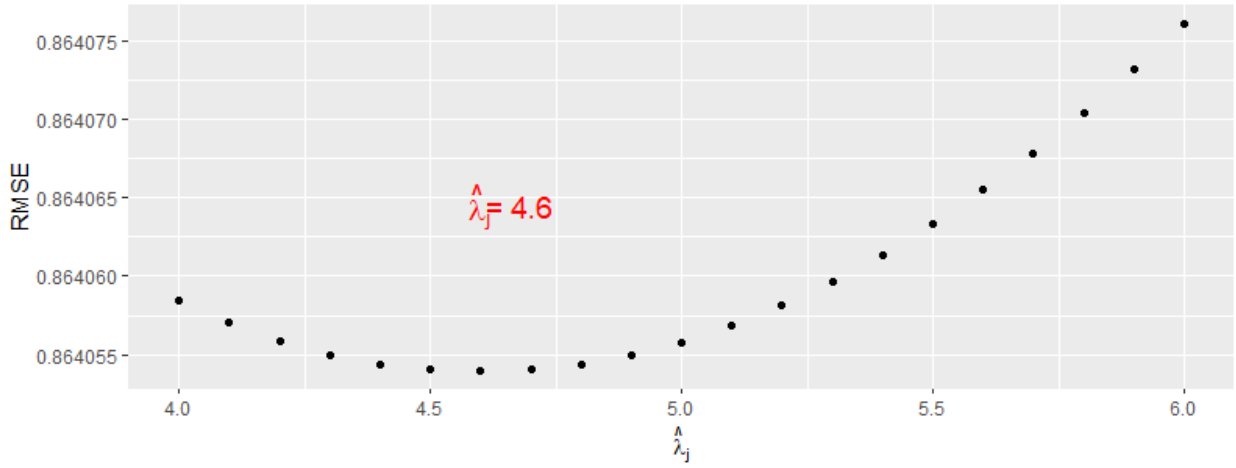
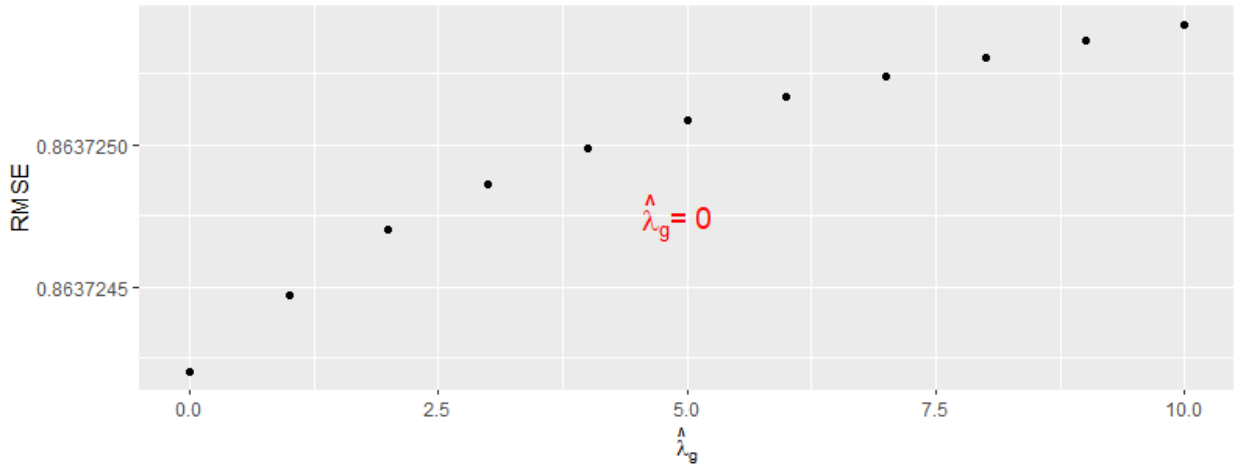


Figure 7: Optimal Value of $\hat{\lambda}_g$ that minimizes the RMSE for $\hat{e}_{ij} = r_{ij} - \bar{b} - \hat{b}_i - \hat{b}_j - \hat{b}_g$



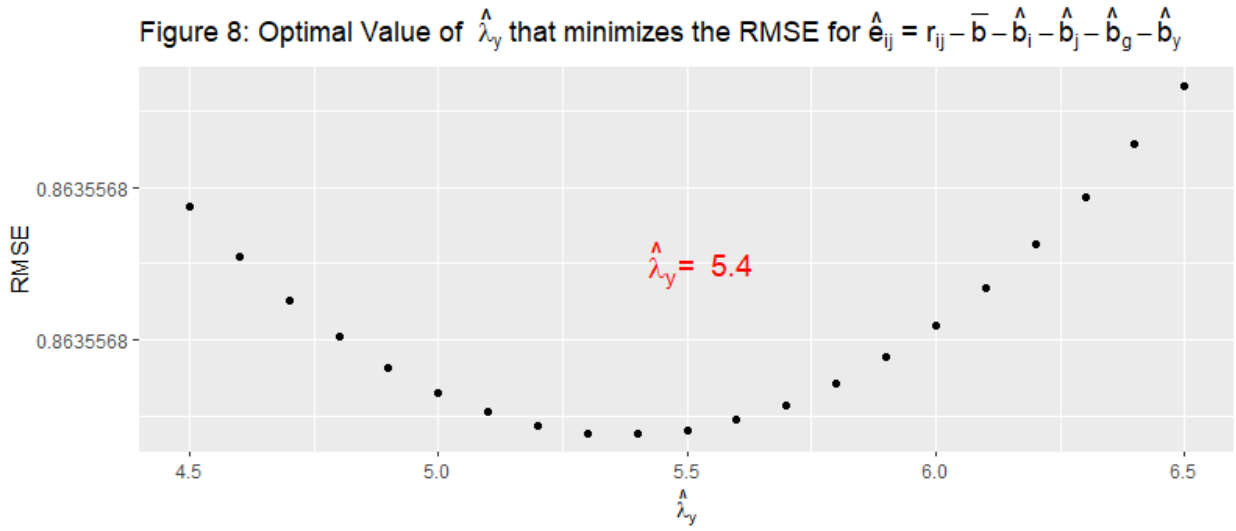


Table 11
RMSE for Each Model Using edxtest Dataset

model	RMSE
The Mean Model	1.0600537
Regularized Model: Movie Effect	0.9429369
Regularized Model: Movie + User Effect	0.8640539
Regularized Model: Movie + User + Genre Effect	0.8637242
Regularized Model: Movie + User + Genre + MovieYear Effect	0.8635568

Adding the individual effects led to a reduction in the RMSE (Table 11). The largest reduction occurred for the movie and user effect. After applying the final model to the validation datatset it produced an RMSE of 0.86464, meeting our target of less than 0.86490.

Section 5: Conclusion

The project's objective was to predict movie ratings in the validation dataset from a model built using the edx dataset. To achieve this the impact of movies, users, genres and movie release year as predictors on ratings was examined. The final model produced an RMSE of 0.8646428, which met the target of an RMSE less than 0.86490.

Limitations

Lack of higher computer power (RAM capacity), which limited the ability of utilizing more advanced machine learning techniques.

Future Work

Collecting additional data on the users, for example, the age and gender of the users, to use as additional predictors in developing a model using a subset of the data with this additional information.